The CAPEC Database

T. L. Nielsen, J. Abildskov, P. M. Harper, I. Papaeconomou & R. Gani*

CAPEC, Department of Chemical Engineering

Technical University of Denmark

DK-2800 Lyngby, Denmark

* E-mail: rag@kt.dtu.dk

Abstract

The CAPEC database on measured data is established with the aim to promote greater data exchange in the chemical engineering community. The target properties are pure component properties, mixture properties and special drug solubility data. The database divides pure component properties into primary, secondary and functional properties. Mixture properties are categorised in terms of the number of components in the mixture and the number of phases present. The compounds in the database have been classified based on the functional groups in the compound. This classification makes the CAPEC database a very useful tool, for example, in the development of new property models since properties of chemical similar compounds are easily obtained. A program with efficient search and retrieval functions of properties has been developed.

Introduction

Measured data are essential to process knowledge, process modelling and to selection of strategies for solving process design problems. They also form the basis of property models. Indeed, data is of key importance, when performing almost any task in Computer-Aided Process Engineering (CAPE). Major advances made in property estimation in the last decade, stem from our ability to comprehensively assess the accuracy and reliability of various model formulations, based on access to even greater numbers of measured data, obtained at even wider ranges of conditions. Recent advances in theory, experiment and molecular simulation have also expanded the options for property determinations, for use in process design. Diverse options for making computations from molecular-level principles[1] are currently being pursued with high-speed computers. Such efforts will probably provide increased guidance about systems at extreme conditions, reaction kinetics details, complex molecules etc.. Thus, future computer-aided process and product engineers will be able to utilise a greater richness of property model options. Yet, paying attention to measurements will probably continue to be important: In the past, reconciliation of new ideas with measurements appears to have formed the basis of most successful innovations. A trend in recent years has been that corporations sell subscriptions to large collections of data that are updated as new data appear. Such subscriptions are expensive and inflexible for the needs of educational institutions or small industrial engineering R & D groups. The CAPEC database project was initiated in 1998 with the objective to establish a flexible, easy to use and easy to maintain, database on measured properties on pure components and mixtures. In recent years, the "WebBook" of NIST[2] and Camsoft[3] "Chemfinder" on

pure component data have initiated data exchange facilitations by providing web-access to their databases.

Classification of Properties

Pure Component Properties

The database includes around 13000 compounds. When available, following pure component properties are given: Acentric factor, critical temperature, critical pressure, critical volume, critical compressibility, melting point, boiling point, triple point temperature and pressure, boiling point at specified pressure, liquid volume at 298K, ideal gas enthalpy at 298K, ideal gas Gibbs energy at 298K, ideal gas entropy at 298K, density, solubility parameters, van der Waals surface area and volume, radius of gyration, dipole moment, octanol/water partition coefficient, refractive index, molecular refraction, heat of fusion, heat of combustion and flash point temperature, dielectric constant. The properties are divided in primary, secondary and functional properties. Primary properties, such as critical temperature and normal boiling point, depend on molecular structure only. Properties such as the ideal gas enthalpy at 298.15K, liquid density at 298.15K or boiling point at a specified pressure (reduced pressure boiling point) are also classified as primary or secondary properties, since the intensive variables are fixed. By tradition, the influence of pressure on liquid densities is neglected here. The characteristic feature of primary and secondary properties is that they are accurately specified by a single-value constant. Functional properties, on the other hand, depend upon temperature, pressure or both. Functional properties are usually represented by a parameterised mathematical correlation along with numerical parameter values that mimic

the property behaviour over a range of conditions. For example, vapour pressure is commonly represented by a set of Antoine parameters instead of raw (T, P^{sat}) data points. Therefore, functional properties are described as numerical values of correlation parameters that have been fitted to the measured data.

Mixture Properties

The mixture properties have been divided into two main categories, namely binary and ternary data. Experimental data for quaternary or higher are rare and are therefore not considered in the first version of the database. Binary data have been collected on VLE, LLE, SLE, infinite dilution activity coefficients, heats of mixing, partial molar heats of mixing at infinite dilution, excess Gibbs energies, Henry's law constants and mutual solubilities. Collected ternary mixture data comprise VLE, LLE, SLE, VLLE, heats of mixing and binodal data. Around 41.000 binary and 10.000 ternary data points have been collected and figure 1 shows how the data are distributed in the different categories. It is noted that VLE data is the largest class of data in the database for both binary and ternary data.

The mixture data were carefully evaluated for errors such as obvious outliners and if errors were encountered the data were cross-checked with the original reference. Perhaps unexpectedly, this procedure often revealed that the errors could be traced back to the original data. It is clear that this is not a test for thermodynamic consistency, but merely a check for obvious errors such as miss types when translating the data from paper to electronic version.

Special Solubility Data

This class of data covers solubilities of drugs such as steroids, penicillins and amino acids. The data have been divided into seven categories: Steroids, amines, benzene halogenated, purine, penicillins and other. This type of data is found in pharmaceutical journals, as not many traditional engineering journals publish such data to a large extent. Presently, the drug solubility database contains around 1500 data points covering 233 solutes and 90 solvents. As shown by figure 2 most of the solubility data are found for the penicillins and steroids categories. In addition to the solubility data a solvent list with solubility indicators (decomposes, miscible, insoluble, slightly soluble, soluble and very soluble) have been constructed for most of the 13000 compounds in the database. For instance, figure 3 shows the distribution of solutes in the 6 categories when water is the selected solvent.

Classification of Compounds

Besides the classification of properties the compounds in the database are classified in nine main categories such as polar compounds, non-associating compounds, electrolytes, steroids. Each main category is divided into several subcategories to gain further information about the functional groups in the compound. For instance, "Esters" and "Nitro" are two subcategories in "Polar Non-Associating Compounds" and "Alcohols" and "Amides" are two subcategories in "Polar Associating Compounds". An overview of the nine different main categories is given in figure 4, where for illustration purposes the subcategories of "Polar Associating Compounds" are shown. This classification of compounds facilitates easy and efficient searching and retrieval of the properties of special families of compounds. Table 1 shows an example of the classification of

2–(Hydroxyimino) propanol oxime and 1,2–benzenedicarboxamide. Based on the structure and functional oxime -groups (=N-OH), 2–(Hydroxyimino) propanol oxime is classified as a "Polar Associating Compound" and subcategorised as an "Oxime" compound. 1,2–benzenedicarboxamide is also classified as an "Polar Associating Compound" but is subcategorised as "Amides" due to the functional amide groups. Figure 5 shows how the compounds in the database distribute in the nine main categories. It is seen that more than one third of the 13000 compounds are multifunctional-grouped compounds. This underlines the necessity of research on property models that are able to handle multifunctional systems.

Database Structure

A database engine has been made to browse/search and retrieve data that are stored in a Microsoft Access database. An overview of the program structure is shown in figure 6. From the main interface one can access the different data categories e.g. pure component properties can be viewed by entering the "Pure Properties" section and mixture properties can be viewed by entering the "Mixture Properties" section. The search for special solubility data is performed in the "Solute Solubility" section, where a list of solvents and the solubility data are shown when a solute is specified. If entering the "Solvent Properties" a list of solvents and the solubility indicators are shown for a specified solute. The last section is the "Solute/Solvents search", where dynamic queries can be setup to search for solvent or solute candidates that satisfy a certain physico-chemical property behavior. For example, the search for a solvent with a melting point greater than 350K and a solubility parameter between 18.0$MPa^{½}$ and 20.0$MPa^{½}$ can

easily be accomplished. In this example naphthalene is one of many possible solvent candidates with a melting point at 353.35K and a solubility parameter 19.45 $MPa^{½}$.

The CAPEC database will allow the user to add his/her own data to the original database. It is also possible to delete/change the data added by the user but not the original data belonging to the CAPEC database. In this way, all users of CAPEC database will share a common database while at the same time, each user will have their own in-house/confidential data for their own use.

Discussion and Conclusion

The CAPEC database project has been initiated and a significant amount of data has been collected including pure component properties, mixture properties and special drug solubility data. The database will continue to grow as new data appear in the literature and especially the mixture properties are the target for future growth. The classification of compounds is very useful in the development of property models such as Constantinou & Gani's method for the prediction of pure component properties[4] or UNIFAC[5] and ASOG[6] methods for the prediction of liquid activity coefficients. With the appropriate tools (currently under development at CAPEC) these methods could easily be extended to handle new families of compounds/mixtures. Also, special group contribution tables could be generated to handle special families of compounds/mixture to increase the accuracy of the predictions compared to the predictions with the original group contribution tables. These are just some of many examples of the usefulness of the classification of the compounds. Finally, the CAPEC database and the database engine can

easily be accessed and used by external programs such as process simulators and other CAPE programs.

Acknowledgement

The authors wish to thank Mitsubishi Chemical, Japan, and Union Carbide, USA, for finical support. Also thanks to Jakob M. Harper and the master students at CAPEC for contributing to the development of the CAPEC database.

References

(1) Gubbins, Keith E., The Future of Thermodynamics, *Chem. Eng. Progress* Feb.1989, 38-49.

(2) NIST Chemistry Webbook can be found at http://webbook.nist.gov/

(3) Chemfinder can be found at http://www.chemfinder.com/

(4) Constantinou, L.; Gani, R., A New Group-Contribution Method for the Estimation of Properties of Pure Compounds, *AIChE J.* 1994, 1697-1710.

(5) Fredenslund, A.; Jones, R. L.; Prausnitz, J. M., Group-Contribution Estimation of Activity Coefficients in Nonideal Liquid Mixtures, *AIChE J.* 1975, vol 21, 1086-1099.

(6) Palmer, David A., Predicting Equilibrium Relationships for Maverick Mixtures, *Chemical Engineering* June 1975, 80-85.

Table of Figures

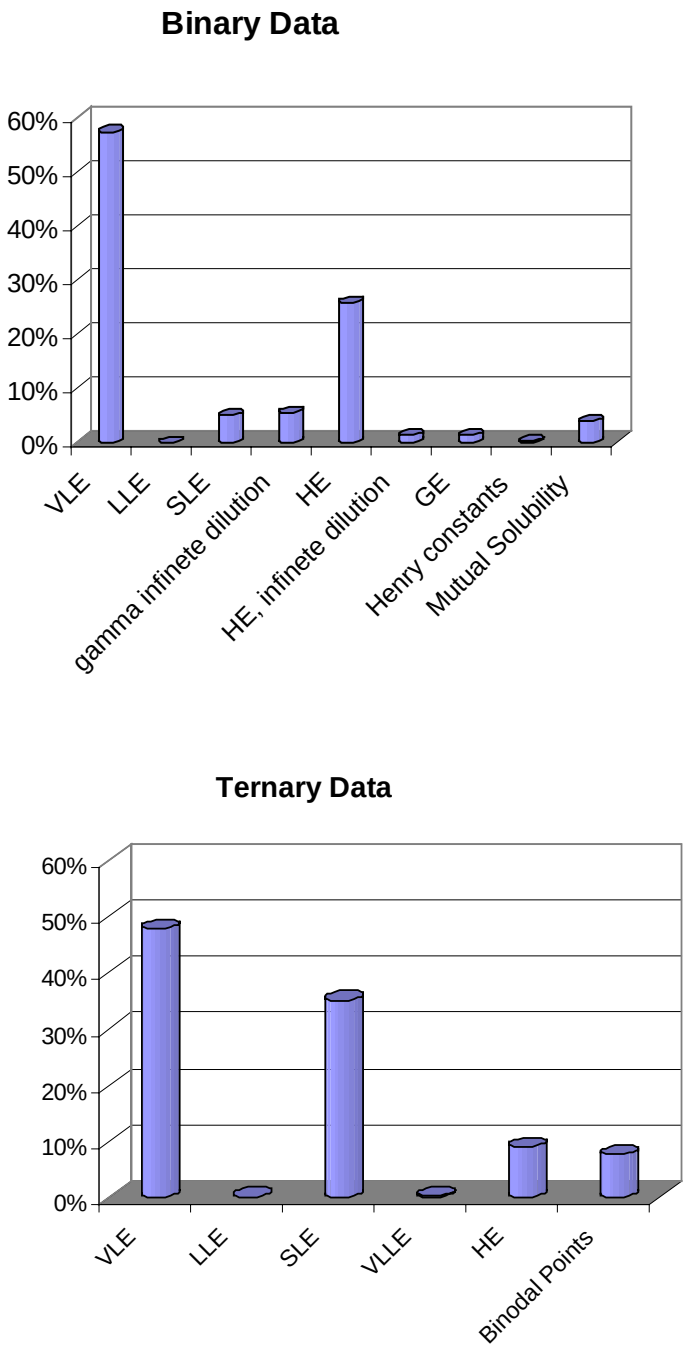


Figure 1: Distribution of collected mixture data among the different categories of data.

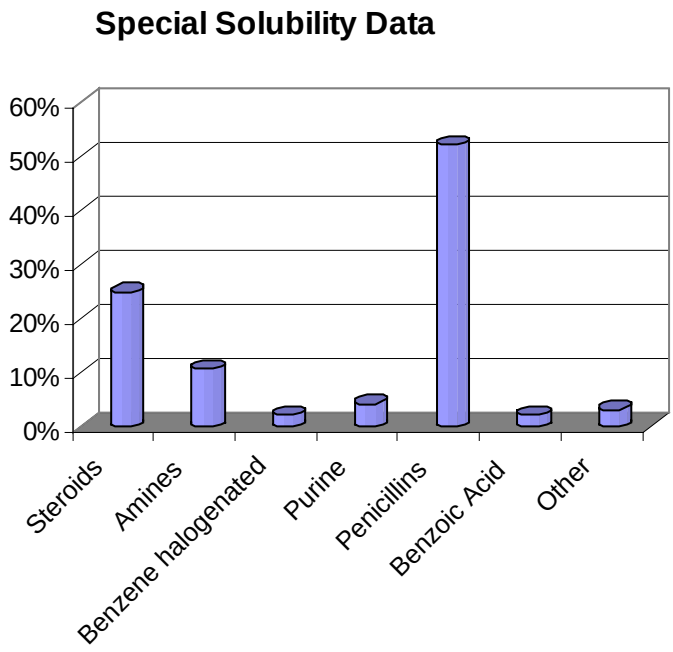


Figure 2: Distribution of special solubility data in the 7 main categories.

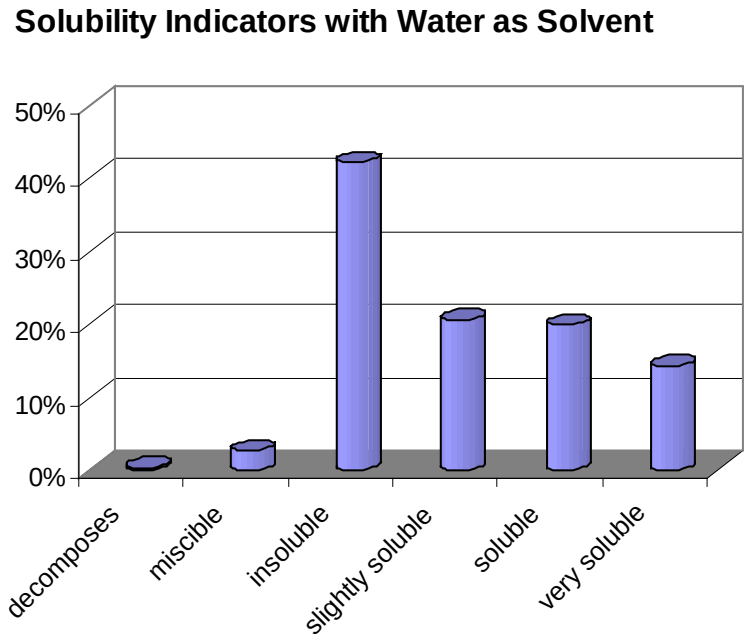


Figure 3: Distribution of solutes with respect to water as solvent.

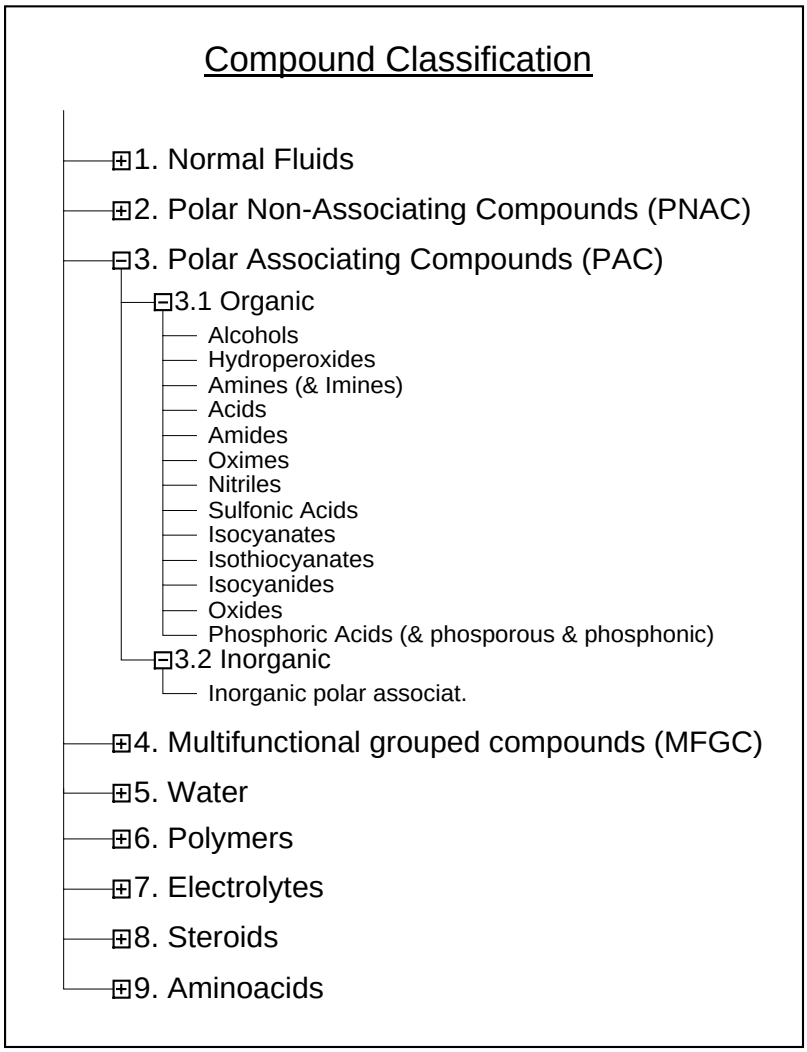


Figure 4: The 9 main categories for the classification of compounds.

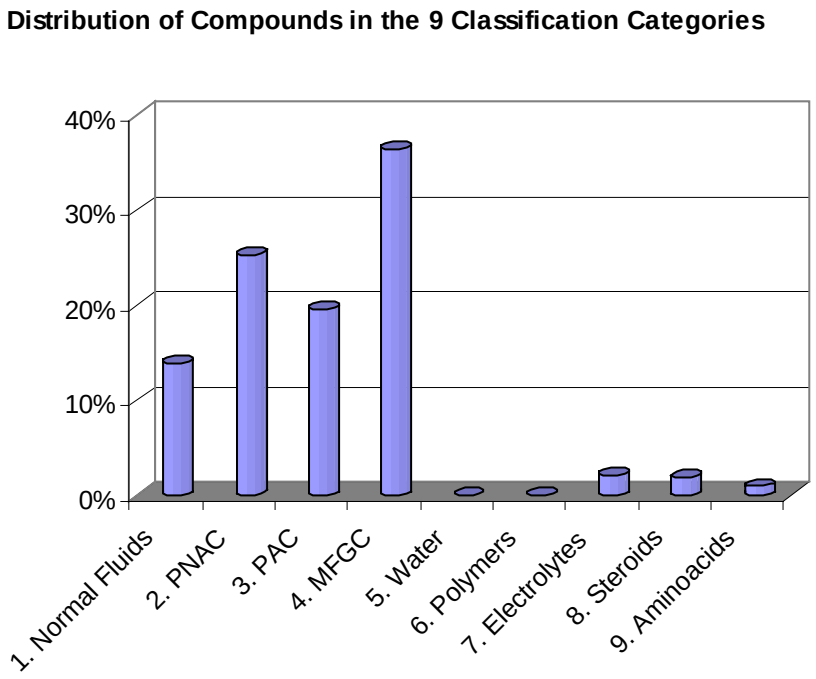


Figure 5: Distribution of the compounds in the CAPEC database in the 9 main categories.

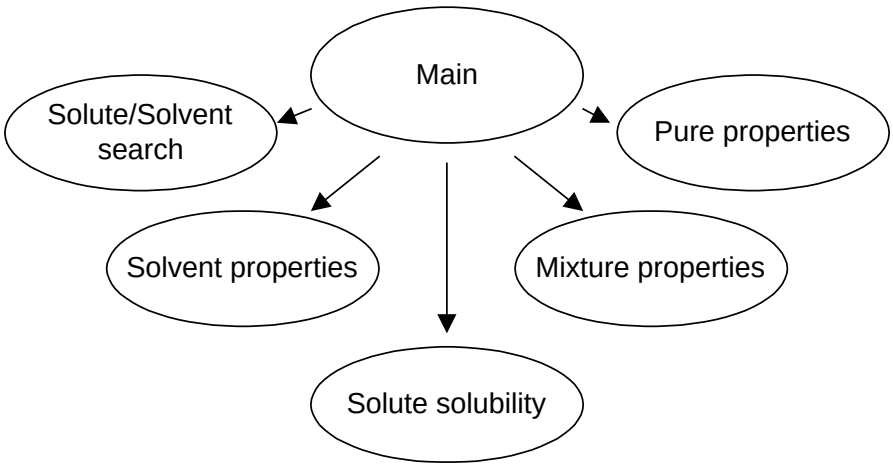
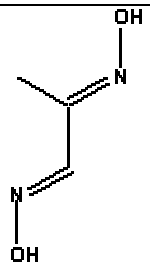
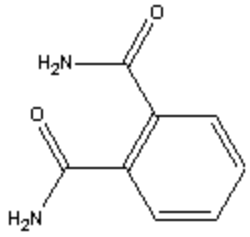


Figure 6: The CAPEC database structure.

Table 1: Classification of 2−hydroxyimino propanol oxime and 1,2−Benzenedicarboxamide.

	
Name: 2-(Hydroxyimino) propanol oxime Classified: "Polar Associating Comp." Subcategory: "Oximes"	Name: 1,2-Benzenedicarboxamide Classified: "Polar Associating Comp." Subcategory: "Amides"